Highly-Smooth Zero-th Order Online Optimization Vianney Perchet

Francis Bach, Vianney Perchet

JMLR: Workshop and Conference Proceedings vol 49:1–27, 2016

Highly-Smooth Zero-th Order Online Optimization

Francis Bach FRANCIS.BACH@ENS.FR
INRIA & Département d'Informatique de l'Ecole Normale Supérieure

Vianney Perchet VIANNEY.PERCHET@NORMALESUP.ORG
CREST - ENSAE

Abstract

The minimization of convex functions which are only available through partial and noisy information is a key methodological problem in many disciplines. In this paper we consider convex optimization with noisy zero-th order information, that is noisy function evaluations at any desired point. We focus on problems with high degrees of smoothness, such as logistic regression. We show that as opposed to gradient-based algorithms, high-order smoothness may be used to improve estimation rates, with a precise dependence of our upper-bounds on the degree of smoothness. In particular, we show that for infinitely differentiable functions, we recover the same dependence on sample size as gradient-based algorithms, with an extra dimension-dependent factor. This is done for both convex and strongly-convex functions, with finite horizon and anytime algorithms. Finally, we also recover similar results in the online optimization setting.

Keywords: Online learning, Optimization, Smoothness

1. Introduction

The minimization of convex functions which are only available through partial and noisy information is a key methodological problem in many disciplines. When first-order information, such as gradients, is available, many algorithms and analysis have been proposed (see, e.g., Shalev-Shwartz, 2011, and references therein), taking the form of stochastic gradient descent (Robbins and Monro, 1951), online mirror descent (Lan et al., 2012), dual averaging (Xiao, 2010) or even variants of ellipsoid methods (Nemirovski and Yudin, 1983; Agarwal et al., 2013). Strong convexity has emerged as an important property characterizing the performance of these algorithms, with optimal convergence rates of $O(1/n)$ after n iterations for strongly-convex problems, and only $O(1/\sqrt{n})$ for convex problems.

However, smoothness can typically only improve constants (Lan, 2012), with the stochastic part of the generalization performance having the same scalings than in the non-smooth case. Apart for quadratic functions or logistic regression where the rates may be improved (Bach and Moulines, 2013; Shamir, 2013; Hazan et al., 2014), the boundedness of high-order derivatives is typically not advantageous.

In this paper, we consider situations where only noisy function values are available, originating from derivative-free optimization (Spall, 2005) and with increased received attention (see, e.g., Bubeck and Cesa-Bianchi, 2012, and references therein). This is also the core assumption in the online learning class of problems known as "bandit" (even though our setup is a bit different, and we obtain faster rates than in bandit optimization).

Again, strong convexity has emerged as a key property (Hazan and Levy, 2014). Following Polyak and Tsybakov (1990), Dippon (2003) or Saha and Tewari (2011) (for the traditional concept of smoothness) we show that in the large variety of online settings, high-order smoothness, namely the boundedness of high-order derivatives, may be used, with the extreme case of infinitely differentiable functions, for which the rates attain the ones for first-order oracles.

More precisely, throughout this paper, we consider a sequence of convex functions $f_n : \mathbb{R}^d \to \mathbb{R}$, $n \geq 1$ and a convex constraint set $K \subset \mathbb{R}^d$ with non-empty interior. The objectives are to output a sequence of a sequence of points $\{x_n\}_{n=0,\dots,N} \in K$ and of queries $\{y_n\}_{n=1,\dots,N} \in \mathbb{R}^d$ to a noisy zero-th order oracle, in order to minimize one of the following criteria:

– *Stochastic optimization*: All functions f_n are equal to f, and the goal is to minimize

$$f(x_N) - \inf_{x \in K} f(x)$$

for the final point $x_N \in K$.

– *Online optimization*: The criterion to optimize, usually referred to as the "regret", is

$$\frac{1}{N} \sum_{n=1}^{N} f_n(x_{n-1}) - \inf_{x \in K} \frac{1}{N} \sum_{n=1}^{N} f_n(x).$$

We immediately emphasize here that a bound valid for online optimization immediately transfers into into a bound for stochastic optimization with the choice $x_N = \frac{1}{N} \sum_{m=0}^{N-1} x_m$.

– *Bandit learning*: this setting is similar to the online optimization case, except that the evaluation point must be equal to the query point, i.e., $y_{n+1} = x_n$ for all n.

Formally, the timing of the optimization scheme is the following. The algorithm first outputs $x_0 \in K$ and queries $y_1 \in \mathbb{R}^d$. After getting $f_1(y_1) + \varepsilon_1 \in \mathbb{R}$ as a feedback (where $\varepsilon_1 \in \mathbb{R}$ is some noise), it outputs $x_1 \in K$ and queries $y_2 \in \mathbb{R}^d$, gets $f_2(y_2) + \varepsilon_2 \in \mathbb{R}$ as feedbacks, etc. Formally, let $\mathcal{F}_{n-1}$ be the σ-field generated by $\{x_0, x_1, y_1, \varepsilon_1, \dots, x_{n-1}, y_{n-1}, \varepsilon_{n-1}\}$. Then x_n and y_n are random variables adapted to $\mathcal{F}_{n-1}$ and ε_n is adapted to $\mathcal{F}_n$.

For simplicity we assume that the noise is independent in the sense that the distributions of ε_n conditionally to $\mathcal{H}_n$ are independent but we do not assume that the noise is identically distributed (as the distribution may depend on y_{n-1}, which is key for online supervised learning). Moreover, we assume that the noise has bounded variance σ^2 that is not necessarily known in advance (improved bounds would be obtained if we allow dependency of algorithms in that term). Note that martingale assumptions common in stochastic approximation (Kushner and Yin, 2003) could be used instead of conditional independence.

Motivating examples for the optimization case are (a) simple additive noise on f, or (b) $f_n(x) = \mathbb{E}_a g(a, x)$ and $\varepsilon_n = g(a_n, x) - \mathbb{E}_a g(a, x)$ for a_n a random variable, which corresponds to online supervised learning where a_n represents the data received at time n.

We shall also consider the case where we essentially query twice the same functions before outputting a new point x_{n+1}; we stress out here that the two feedbacks are two noisy evaluations where *the noises are independent*, as opposed to Agarwal et al. (2010); Duchi et al. (2013). As a consequence, the classical optimization setup remains identical except that we make $2N$ queries

	Stochastic Constrained & Unconstrained	Online Constrained & Unconstrained
Convex $\beta = 2$	$(\frac{d^2}{N})^{\frac{1}{3}}$	$(\frac{d^2}{N})^{\frac{1}{3}}$
$\beta > 2$	$(\frac{d^2}{N})^{\frac{\beta-1}{2\beta}}$	$(\frac{d^2}{N})^{\frac{\beta-1}{2\beta}}$
μ-stg convex $\beta = 2$	$\sqrt{\frac{d^2}{\mu N}}$	$\sqrt{\frac{d^2}{\mu N}}$
$\beta > 2$	$(\frac{d^2}{\mu N})^{\frac{\beta-1}{\beta+1}}$	$(\frac{d^2}{\mu N})^{\frac{\beta-1}{\beta+1}}$
(asymptotic)	$\frac{1}{\mu^2}(\frac{d^2}{N})^{\frac{\beta}{\beta+1}}$	

Figure 1: Summary of the principal rates of convergence achieved by our algorithms for stochastic or online optimization. The bounds in the last asymptotic regime are only true when N is large enough and are only valid for stochastic optimization.

instead of N, thus rates of convergence are independent of this trick. As a consequence, it only makes a difference in the online optimization setup, where we now need to assume that the same function is observed twice in a row.

We introduce this two-point setting as it allows us to consider the case where the constraint set is the whole space $\mathbb{R}^d$. Moreover, the algorithms do not need to perform a projection at each step and rates of convergence are independent of the maximal value of the loss functions (which should not appear as the problem in translation invariant). Note that (a) this unconstrained setting is common in smooth optimization, and (b) that our proof technique can extend to composite optimization where a non-smooth term is added with its proximal operator (Xiao, 2010; Hu et al., 2009). On the other hand, when the constraint set is a compact convex subset, of diameter denoted by $R > 0$, then we shall use a classical "one-point" algorithm that queries each f_n only once.

We shall provide algorithms and explicit rates of convergence for all the following cases

i) Unconstrained ($K = \mathbb{R}^d$) vs. constrained optimization (K is compact convex).

ii) Convex vs. μ-strongly convex mappings.

iii) Stochastic optimization vs. online optimization.

Maybe surprisingly, as shown in Figure 1, rates of convergence are actually independent of the unconstrained/constrained setting and on the stochastic vs. online case, at least when f_n are Lipschitz-continuous which is a required setup for online optimization. We emphasize here that the asymptotic dependencies in N and d are exact, i.e., no logarithmic terms are hidden.

Note that we do not consider here the bandit setting that imposes that $x_{n+1} = y_n$. This can be deduced from Figure 1 as the rate for strongly convex functions would violate the lower bound of Shamir (2013) for bandit learning.

1.1. Smoothness assumption

We shall assume that all mappings in question are defined on $\mathbb{R}^d$ and almost surely $(\beta-1)$-times differentiable and that for all $\|v\|_2 = 1$, and $x, y \in \mathbb{R}^d$, then

$$\|f^{(\beta-1)}(x)v^{\beta-1} - f^{(\beta-1)}(y)v^{\beta-1}\|_2 \leq M_\beta^\beta \|x-y\|_2, \tag{1}$$

where we define

$$f^{(m)}(x)v^m = \sum_{m_1+\cdots+m_d=m} \frac{\partial^m f}{\partial^{m_1}x_1 \cdots \partial^{m_d}x_d} v_1^{m_1}\cdots v_d^{m_d}$$

as the m-th term in the Taylor expansion of f. We refer to such functions as *β-th order smooth* functions. Note that a stronger assumption is that f is β-times differentiable with a uniform bound

$$\sup_{x\in\mathbb{R}^d} \sup_{\|v\|\leq 1} |f^{(\beta)}(x)v^\beta| \leqslant M_\beta^\beta. \tag{2}$$

These notions extends the traditional smoothness, which corresponds to $\beta = 2$ (Nesterov, 2004). Notice that this implies that for all x, y (as a consequence of Taylor expansions with integral remainder):

$$\left| f(y) - \sum_{|m|\leq\beta-1} \frac{1}{m!} f^{(m)}(x)(y-x)^m \right| \leq \frac{M_\beta^\beta}{\beta!}\|y-x\|^\beta. \tag{3}$$

We emphasize the fact that high-order smoothness, in the sense defined above, implies lower order smoothness only if mappings are defined on a compact set. If a mapping is defined on the whole space, then it can be second order smooth without being first order smooth, such as any non trivial quadratic function.

We now mention the following lemma that relates the different degrees of smoothness of f.

Lemma 1 *Let $f : K \to \mathbb{R}$ be a continuous mapping that is β_1-smooth and β_2-smooth, with the associate constants M_{β_1} and M_{β_2}, where $\beta_1 < \beta_2$. Then f is β-smooth for all $\beta \in [\beta_1, \beta_2]$ and there exist a sequence of weights α_β, for all $\beta \in [\beta_1, \beta_2]$, independent of M_{β_1} and M_{β_2} such that*

$$\alpha_\beta M_\beta^\beta \leq 2(\alpha_{\beta_1} M_{\beta_1}^{\beta_1})^{\frac{\beta_2-\beta}{\beta_2-\beta_1}} (\alpha_{\beta_2} M_{\beta_2}^{\beta_2})^{\frac{\beta-\beta 1}{\beta_2-\beta_1}}$$

In particular,

i) if K is compact then f is bounded (i.e., 0-smooth). As a consequence, β-smoothness immediately entails that f is Lipschitz and 2-smooth.

ii) If f is Lipschitz and β-smooth (for $\beta \geq 2$), then f is 2-smooth.

From now on, we shall assume that all mappings f_n are β-smooth, for some $\beta \geq 2$, with a common associated constant M_β which is known (which typically holds in many settings, see next example). In online unconstrained optimization, we will also impose that f_n is Lipschitz (again, this is automatic when K is compact).

Special case: logistic regression. If $f(x) = \mathbb{E}_a \log(1 + \exp(-a^\top x))$ for a certain random vector $a \in \mathbb{R}^d$ which is uniformly bounded by R, then we consider $\varepsilon_n = \log(1 + \exp(-a_n^\top x)) - \mathbb{E}_a \log(1 + \exp(-a^\top x))$ for a sample a_n. This is online logistic regression, for which the constant M_β^β may be chosen to be equal to $\frac{1}{4}(\beta-1)!R^\beta$ (Kakade et al., 2009), which is such that $M_\beta \leq \beta R$. Note that such a setting should extend to all generalized linear models. Moreover, we use a property of logistic regression which is different than self-concordance (Bach, 2010), which bounds the third derivatives by the second derivative; it would be interesting to see if the two analyses can be combined.

1.2. Related work

As already mentioned, there is a huge (and actually still increasing) literature on stochastic optimization with zero-th order feeback and/or on convex bandits problem. We also investigate here the online optimization setup, an "intermediate framework" where the sequence of mappings f_n can evolve adversarially but, as in optimization, the loss might be evaluated at another point than the query sent to the oracle.

We emphasize these differences between set-ups as the complexity of stochastic zero-th order optimization and the convex bandit problem have been widely studied recently (Recht et al., 2012; Shamir, 2013). It has been observed that minimax rates of convergence in bandit problems and stochastic optimization might differ, which is not the case in our setting for our upper-bounds (one can therefore conclude that the complexity of convex bandits is not hidden in the evolving sequence of loss functions, but more importantly on the constraint that the query point is where the loss is evaluated).

Moreover, it has also been shown by Recht et al. (2012); Shamir (2013) that the slow rates of $\sqrt{d^2/n}$ are minimax optimal for stochastic optimization or convex bandits. The optimal rates of $\sqrt{1/n}$ have been obtained (Nemirovski and Yudin, 1983; Liang et al., 2014) but without the explicit dependency in the dimension d; moreover, those techniques cannot be used in online optimization. The lower bound in $\sqrt{d^2/n}$ holds even if the mappings are highly regular, as quadratic and strongly-convex (Shamir, 2013). However, in that case, the optimization error decreases as d^2/n; see also Hazan et al. (2014) for a similar result on logistic regression. This result[1] can be interpreted as an extreme case of our regularity assumptions, i.e., when $\beta = +\infty$ or $M_3 = 0$. As a consequence, we somehow interpolate between the well studied extreme problems in online learning with either smooth or quadratic mappings.

The intermediate framework between smooth and quadratic (or mappings infinitely differentiable) has also been studied by Fabian (1967), Chen (1988) and Polyak and Tsybakov (1990) where the focus was stochastic optimization with the objective of bounding the error in the argument and not in function evaluation. Fabian (1967) obtained an algorithm such that the distance to the maximum is of the order of $N^{-\frac{\beta-1}{2\beta}}$ which is optimal (Chen, 1988). In the case of strongly-convex mappings, this has been improved by Polyak and Tsybakov (1990) to $N^{-\frac{\beta-1}{\beta}}$ which is also optimal. Our set-up is more general (as we consider also online learning, function evaluations) and we recover the aforementioned results as a byproduct of ours, with a novel non-asymptotic analysis with an explicit dependencies in the dimension and parameters of smoothness and strong convexity.

1. Actually, the quadratic case is very particular as we could show that one can query points arbitrarily away from the origin to reduce variance.

2. Smoothing Lemma

Our analysis relies on a novel single stochastic approximation lemma, which combines ideas from Nemirovski and Yudin (1983); Nesterov (2011) and Polyak and Tsybakov (1990). Let f be a convex function defined on $\mathbb{R}^d$.

Expectation of random function evaluations around a point. Given positive scalars $\delta, r > 0$, we consider sampling the value $f(x + r\delta u)$ around x, for u uniformly distributed in the unit *sphere* for the Euclidean norm. As shown by Nemirovski and Yudin (1983), the expectation of the vector $f(x + r\delta u)u$ is equal to $d/(\delta r)$ times the gradient of a function which is an approximation of f, that is, $x \mapsto \mathbb{E}_{\|v\|_2 \leqslant 1} f(x + \delta r v)$, where v is now sampled uniformly from the unit *ball*. This simple result is a consequence of Stokes' theorem[2] Thus the expectation of function evaluations at random points around x is the gradient of a certain function. This is a key property which is used by most non-asymptotic analyses (Flaxman et al., 2005) of zero-th order optimization.

High-order smoothness and gradient evaluation. As shown by Polyak and Tsybakov (1990) in one dimension (and then generalized to partial derivatives), if we now sample independently r from the uniform distribution in $[-1, 1]$, and we consider a function $k(r)$ such that $\mathbb{E}_r r k(r) = 1$ and $\mathbb{E}_r r^k k(r) = 0$ for k odd between 3 and β, then $\frac{1}{\delta} f(x + \delta r) k(r)$ is a good approximation of the derivative of f at x, with an expectation (with respect to r) which is equal to $f'(x)$ up to terms of order $\delta^{\beta - 1}$ if f is β-th order smooth.

In the following lemma, we combine these two ideas (see proof in Appendix A.2):

Lemma 2 *Let $f : \mathbb{R}^d \to \mathbb{R}$ a convex function. Define*

$$\hat{f}_\delta(x) = \mathbb{E}_r \mathbb{E}_{\|v\| \leq 1} f(x + r\delta v) r k(r),$$

where the expectation is taken with respect to the uniform distribution on the unit ball for v, and $r \in \mathbb{R}$ is independent from v, with uniform distribution in $[-1, 1]$, and $k(r)$ is such that $\mathbb{E}_r r k(r) = 1$ and $\mathbb{E}_r r^k k(r) = 0$ for k odd between 3 and β. Then, $\hat{f}_\delta$ is differentiable and for any $x \in \mathbb{R}^d$,

$$\hat{f}'_\delta(x) = \frac{d}{\delta} \mathbb{E}_r \mathbb{E}_{\|u\|_2 = 1} \big[f(x + \delta r u) k(r) u \big]. \tag{4}$$

Moreover, we have the approximation bounds (the second being valid if f is differentiable):

$$\begin{aligned} \big| \hat{f}_\delta(x) - f(x) \big| &\leq \frac{M_\beta^\beta}{\beta!} \delta^\beta \Big(\mathbb{E}_r \big| k(r) r^{\beta + 1} \big| \Big), \\ \| \hat{f}'_\delta - f'(x) \| &\leq \frac{M_\beta^\beta}{(\beta - 1)!} \delta^{\beta - 1} \Big(\mathbb{E}_r \big| k(r) r^\beta \big| \Big). \end{aligned}$$

Choice of $k(r)$. Following Polyak and Tsybakov (1990), we consider r uniformly distributed in $[-1, 1]$. For $\beta \in \{1, 2\}$, we may take $k(r) = 3r$, for which we have $\mathbb{E} r k(r) = \frac{1}{2} \int_{-1}^1 3r^2 dr = 1$.

Consider orthonormal polynomials $p_m(\cdot)$ for the distribution on r, i.e., such that $\mathbb{E}_r p_m p_{m'} = 0$ for $m \neq m'$, $\mathbb{E}_r p_m^2 = 1$ and $p_0(\cdot), \dots, p_s(\cdot)$ spans the vector space of polynomials of degree less or equal than s, for all $s \in \mathbb{N}$.

2. Without loss of generality, we may consider $r\delta = 1$ and $\mathbb{B}$ the unit ball; then the gradient of $x \mapsto \mathbb{E}_{\|v\|_2 \leqslant 1} f(x + v)$ is $\frac{1}{\mathrm{vol}(\mathbb{B})} \int_{\mathbb{B}} f'(x + v) dv = \frac{1}{\mathrm{vol}(\mathbb{B})} \int_{\partial \mathbb{B}} f'(x + u) du$ by Stokes' theorem and because u a normal vector to the unit sphere $\partial \mathbb{B}$ at u. The factor of d comes from the ratio between the volume of the ball and the surface of the sphere.

Then we may choose $k(r) = \sum_{m=0}^{\beta} p'_m(0) p_m(r)$. Indeed, following Polyak and Tsybakov (1990), given $s \in \mathbb{N}$, let $b_0, \dots, b_s$ be the coordinates of r^s in the chosen basis, i.e., $r^s = \sum_{j=0}^{s} b_j p_j(r)$, then $\mathbb{E}_r k(r) r^s = \sum_{j=0}^{s} b_j p'_j(0) = 0$ for $s \neq 1$ and zero for $s \in \{0, 2, \dots, \beta\}$. Note that this is more than we actually need as in Lemma 2, we only need s being odd.

We have, for r uniform in $[-1, 1]$, $p_m(u) = \sqrt{2m+1} L_m(u)$ where L_m is the m-th Legendre polynomial. For example, we have the following values for $\beta \in \{1, 2, 3, 4, 5, 6\}$:

$$\begin{aligned} k_1(r) = k_2(r) &= 3r \\ k_3(r) = k_4(r) &= \frac{15r}{4}(5 - 7r^3) \\ k_5(r) = k_6(r) &= \frac{195r}{64}(99r^4 - 126r^2 + 35). \end{aligned}$$

Bounds. In this paper, we also need the following bounds, which are shown in Appendix A.3 by using properties of Legendre polynomials:

$$\begin{aligned} \mathbb{E}_r |k(r)|^2 &\leq 3\beta^3 \\ \mathbb{E}_r |k(r)|^2 r^2 &\leq 8\beta^2 \\ \mathbb{E}_r |k(r) r^{\beta+1}| &\leq 2\sqrt{2}\beta. \end{aligned}$$

Convexity. With respect to the kernel chosen, $\hat{f}_\delta$ is always convex for $\beta = 2$, because $rk(r)$ is always non-negative. For $\beta \geq 3$, if f is μ-strongly-convex, then $\hat{f}_\delta$ is $\mu/2$-strongly-convex if δ is small enough.

Indeed, by definition of $\hat{f}_\delta$ and by 3-smoothness of f, we obtain that

$$D^2 \hat{f}_\delta(x) = \mathbb{E}_r \mathbb{E}_{\|v\| \leq 1} D^2 f(x + r\delta v) r k(r) \succcurlyeq \mu I_d - \delta M_3^3 \mathbb{E}_r |k(r)| r^2 J_d,$$

where J_d is the matrix whose components are all equal to 1. As a consequence, $\hat{f}_\delta$ is $\mu/2$-strongly-convex as soon as $\delta \leq 16\mu/(d\beta^2 M_3^3)$. Note however that $\hat{f}_\delta$ is not convex in general.

3. Unconstrained Optimization

We recall that $f_n = f$ in this setting and that we chose to make two queries y_{n_-}, y_{n_+} of f before outputting the next point x_n. Of course, *stricto sensu*, one should replace N by $N/2$ in our rates of convergence. For simplicity and consistency in proofs, we chose to keep the formulation as N stages of 2 queries. Moreover, the two independent noises can be combined into a single one.

We thus consider two-point algorithms of the form

$$x_n = x_{n-1} - \gamma_n \frac{d}{2\delta_n} \big[f(x_{n-1} + \delta_n r_n u_n) - f(x_{n-1} - \delta_n r_n u_n) + \varepsilon_n \big] k(r_n) u_n, \tag{5}$$

where γ_n and δ_n are constants that depend on n, u_n is uniform in the unit-sphere, and $k(r_n)$ satisfies the conditions of Lemma 2. We emphasize again that the noise is different at the two evaluations points $y_{n_-} = x_{n-1} - \delta_n r_n u_n$ and $y_{n_+} = x_{n-1} + \delta_n r_n u_n$ and do not cancel by differencing (the random variable ε_n is thus the difference of these two zero-mean independent noises). We define $\bar{x}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} x_k$ as the averaged iterate.

3.1. Convex Mappings

We first consider the case of convex (i.e., not necessarily strongly-convex) mappings. In order to preserve the flow of the paper, we delay the proof to Appendix C.1.

Proposition 3 (Unconstrained, Convex) *Assume f is (a) β-th order smooth with constant M_β, and (b) 2nd-order smooth with constant M_2.*

Consider the algorithm in Eq. (5), with $\gamma_n = \gamma = \frac{1}{24 d^{(\beta-1)/\beta} M_2^2 \beta^2 N^{(\beta+1)/(2\beta)}}$ and $\delta_n = \delta = \frac{\beta d^{1/\beta}}{N^{1/(2\beta)}} (M_\beta^\beta M_2)^{-1/(\beta+1)}$ for $n \in \{1, \dots, N\}$. Then, $\mathbb{E} f(\overline{x}_{N-1}) - f(x^*)$ is less than

$$\Big(\frac{d^2}{N}\Big)^{(\beta-1)/(2\beta)} \left(7\beta M_2 \|x_0 - x_*\| + 3\sigma + (M_\beta/M_2)^{2\beta/(\beta+1)} + \frac{\beta}{N^{1/\beta}} (M_\beta/M_2)^{-\beta/(\beta+1)} \right)^2 .$$

We can make the following observations about this proposition:

- **Dominating term in $\left(\frac{d^2}{N}\right)^{(\beta-1)/(2\beta)}$**: the second term in the bound above is asymptotically negligible when N grows and we recover the same scaling as the one-point estimate late in Section 4, with the same scalings for the step size.
- **Recovering the optimal rate of $\frac{1}{\sqrt{N}}$:** If β is infinite then one can consider $\beta = \log_2(N)/2$ to recover the optimal rate (up to logarithmic factor) since $2\sqrt{N}^{\beta/(\beta+1)} \geq \sqrt{N}$. Formally, the rate of convergence would also depend on $M_{\log_2(N)/2}$ that has to grow slowly; for logistic regression, this term is also logarithmic.

  This rate is also achieved if $M_\beta = 0$, a situation that can occur if f is a polynomial, by taking δ of the order of a constant and γ of the order of $1/\sqrt{N}$.
- **Anytime version**: as shown in Appendix C.1, by using decaying step-sizes, we obtain an anytime result (i.e., a result valid for all $N \in \mathbb{N}$) with an extra factor of $\log(N+1)$.

3.2. Strongly-Convex Mappings

We now consider the case of μ-strongly-convex mappings. We emphasize here that, in the following proposition, fast rates of convergence are achieved with non-uniform averages, i.e., we introduce $\widehat{x}_{n-1} = \frac{2}{n(n+1)} \sum_{k=0}^{n-1} (k+1) x_k$. We again delay the proof to Appendix C.2.

Proposition 4 (Unconstrained, Strongly-convex, 2-smooth) *Assume f is (a) β-th order smooth with constant M_β, and (b) 2nd-order smooth with constant M_2.*

Consider the algorithm in Eq. (5), with $\gamma_n = \frac{1}{\mu n}$ and $\delta_n = \left(\frac{d^2 \beta!}{M_\beta^\beta \mu n}\right)^{1/(\beta+1)}$, for $n \in \{1, \dots, N\}$. Then, $\mathbb{E} f(\widehat{x}_{N-1}) - f(x^*)$ is less than

$$\Big(\frac{d^2 M_\beta^2}{n\mu}\Big)^{(\beta-1)/(\beta+1)} \left(8\beta M_\beta \|x_0 - x_*\| + 4\sigma + 2 + \beta (M_2/M_\beta)^2 \Big(\frac{M_\beta^2}{n\mu}\Big)^{2/(\beta+1)} \right)^2 .$$

We emphasize here that the first bound allows to recover the previous bound for the optimization of a non-strongly-convex mapping f by using the aforementioned scheme to $f + \mu\|\cdot\|^2$ and let μ depend on n. The second bound has the optimal dependency in N but a worse dependency in μ.

4. Constrained Optimization

In this setup, where the constraint set K is compact convex and of diameter R, we use a classical one-point algorithm:

$$x_n = \Pi_K\left(x_{n-1} - \gamma_n \frac{d}{\delta_n}\left[f(x_{n-1} + \delta_n r_n u_n) + \varepsilon_n\right] k(r_n) u_n\right), \tag{6}$$

where the parameter γ_n and δ_n can evolve with time. In particular, we have $y_n = x_{n-1} + \delta_n r_n u_n$.

4.1. Convex Mappings

Again, we begin with the case of convex (i.e., non necessarily strongly-convex) mappings. The proof of the following proposition is delayed to Appendix D.1.

Proposition 5 (Constrained, Convex) *Assume f is β-th order smooth with constant M_β and consider the algorithm in Eq. (6), with $\gamma_n = \frac{R\delta_n}{\sqrt{\beta^3 d\sqrt{n}}}$ and $\delta_n^\beta = \frac{d\sqrt{\beta}(\beta-1)!}{\sqrt{n}M_\beta^\beta}$, for $n \in \{1, \dots, N\}$. Then, $\mathbb{E}f(\overline{x}_N) - f(x^*)$ is less than*

$$25RM_\beta\left(\frac{d^2\beta}{N}\right)^{\frac{\beta-1}{2\beta}}(C_{\delta_1} + \sigma^2 + 1),$$

where C_δ is a uniform bound of f on the δ-neighborhood of K.

We can make the following observations:

- **Anytime algorithm:** The algorithm is independent of N, thus it is anytime, i.e., the above rate holds for all $N \in \mathbb{N}$. Notice also that C_{δ_1} can actually be replaced, asymptotically, by C_0; see the proof in Appendix D.1.
- **Upper-bounding C_δ:** Since the mapping f is bounded on the compact set K and β-smooth, it is necessarily M_1-Lipchitz. Then C_δ is bounded by $C_0 + M_1\delta$;
- **Concerning the unknown quantities (C_δ and σ^2):** The step-sizes do not depend on the unknown quantities C_δ or σ^2. However, if they are known, then the dependency on C_0 and σ^2 can be slightly improved. Similarly, we assumed that the constant M_β was known. If it is not the case, the algorithm still works with the specific choice of $\delta_n^\beta = dR\sqrt{\beta}(\beta-1)!/\sqrt{n}$; the dependency in M_β would be changed from M_β into M_β^β.

4.2. Strongly-Convex Mappings

Similarly to the unconstrained case, we now consider the case of μ-strongly-convex mappings where rates can be improved. As before, we delay the proof of the following proposition to Appendix D.2.

Proposition 6 (Constrained, Strongly-convex) *Assume f is β-th order smooth with constant M_β. Consider the algorithm in Eq. (6), with $\gamma_n = 1/(n\mu)$ and $\delta_n = \left(\frac{d^2\beta\beta!}{n\mu M_\beta^\beta}\right)^{\frac{1}{\beta+1}}$ for $n \in \{1, \dots, N\}$. Then, $\mathbb{E}f(\overline{x}_N) - f(x^*)$ is less than*

$$15\beta^2 M_\beta^{\frac{2\beta}{\beta+1}}\left(\frac{d^2}{\mu N}\right)^{\frac{\beta-1}{\beta+1}}(C_{\delta_1} + \sigma^2 + 1)\,.$$

We emphasize the fact that the algorithm is again independent of N, thus the result is actually anytime.

5. Refined Upper and Lower Bounds

In this section, we consider improved bounds in the smooth case ($\beta = 2$), as well as asymptotic and lower bounds for strongly-convex mappings for all β.

As mentioned at the end of Section 2, if $\beta = 2$ then $\hat{f}_\delta$ is always convex. As a consequence, the analysis of the algorithms can be improved by noting that Eq. (5) and Eq. (6) correspond to an exact stochastic gradient descent of the approximate mapping $\hat{f}_\delta$. We recall that the analysis for $\beta \geq 3$ was based on the fact that Eq. (5) and Eq. (6) correspond to an approximate stochastic gradient descent of f.

The differences between f' and $\hat{f}'_\delta$ is of the order of $\delta^{\beta-1}$ while $\hat{f}_\delta$ is δ^β-close to f (disregarding the other dependencies in the dimension d and smoothing parameter β). As a consequence, when $\beta = 2$, we can replace the error term in $\delta^{\beta-1}$ when approximating gradients by δ^β, as we approximate the value functions. Using this idea, and following the same lines of proof, we obtain the following proposition (see proof in Appendix D.3).

Proposition 7 (The case $\beta = 2$) *Assume that f is 2-smooth, then the algorithms described in Eq. (6) and Eq. (5), with adapted choices of parameters, ensures the following upper-bound on $\mathbb{E}f(x_N) - f(x^\star)$:*

– *for unconstrained optimization of convex mappings,*

$$2\left(\frac{d^2}{N}\right)^{\frac{1}{3}}\left(96M_2^2\|x_0 - x_*\|^2 + \frac{\sigma^2}{10} + 18\right) + \frac{2d^2}{N},$$

– *for unconstrained optimization of strongly-convex mappings,*

$$4\big(2\sigma^2 + 27\big)\sqrt{\frac{d^2M_2^2\log(N)}{N\mu}} + \left(\frac{21d^2M_2^2\log(N)}{N\mu}\right)^{3/2},$$

– *for constrained optimization of convex mappings,*

$$44\left(\frac{d^2M_2^2R^2}{N}\right)^{\frac{1}{3}}(C_{\delta_1} + \sigma^2 + 1),$$

– *for constrained optimization of strongly-convex mappings,*

$$66\sqrt{\frac{d^2M_2^2}{\mu N}}(C_\delta^2 + \sigma^2 + 1).$$

We mention here that if we had just plugged the value $\beta = 2$ in the general propositions, we would have got rates of convergence of the order of $n^{-1/4}$ and $(\mu n)^{-1/3}$, instead of $n^{-1/3}$ and $(\mu n)^{-1/2}$, respectively in the non-strongly and μ-strongly-convex case.

Similarly, we have proved that if f is μ-strongly-convex and δ is small enough, then $\hat{f}_\delta$ is $\mu/2$-strongly-convex. As a consequence, the previous arguments hold and we can, asymptotically, obtain better rates of convergences, as we now show (see proof in Appendix D.4).

Proposition 8 (Asymptotics with strongly-convex mappings)
Assume that f is β-smooth, μ-strongly-convex and globally optimized at $x^\star$ on K. Then the algorithms described in Eq. (6) and Eq. (5), with adapted choices of parameters, ensure the following upper-bound on $\|x_N - x^\star\|$ as soon as N is big enough:

– *for unconstrained optimization of strongly-convex mappings,*

$$\frac{16M_\beta^2}{\mu^2}(2\sigma^2+16)\left(\frac{d^2\log(N+1)}{N}\right)^{\frac{\beta-1}{\beta}} + \frac{48\beta^3 M_2^4}{\mu^2 M_\beta^2}\left(\frac{d^2\log(N+1)}{N}\right)^{\frac{\beta+1}{\beta}},$$

– *for constrained optimization of strongly-convex mappings,*

$$16\beta\left(\frac{d^2}{N}\right)^{\frac{\beta-1}{\beta}}\left(\frac{2eM_\beta M_2}{\mu}\right)^2\left(3C_\delta^2+3\sigma^2+1\right).$$

We recall that from those upper-bounds , we obtain $\mathbb{E}f(x_N) - f(x_*) \leq \frac{M_2^2}{2}\mathbb{E}\|x_N - x_*\|_2^2$.

The proof is delayed to Appendix D.4.

We conclude this section with a lower bound for the optimization of strongly-convex mappings, brought to our attention by O. Shamir and based on techniques from Shamir (2013). This lower bounds matches the lower bound of Polyak and Tsybakov (1990), but it is non-asymptotic, quite simple and one can obtain explicit dependencies in the different parameters. We only sketch it in one dimension, as it contains all the relevant ideas; details can be found in Shamir (2013).

Consider the two mappings

$$f_1(x) = 2\mu x^2 + \alpha g\left(\frac{x}{\theta}\right) \text{ and } f_2(x) = x^2 - \alpha g\left(\frac{x}{\theta}\right), \text{ where } g(y) = \frac{y}{1+y^2},$$

and notice that $f_1(x) = f_2(-x)$, $|g(y)| \leq 1/2$ and $\left|g^{(\beta)}(y)\right| \leq 2^{\beta+1}\beta! \leq (2\beta)^\beta$. As a consequence, it is not difficult to see that $\|f_1 - f_2\|_\infty \leq \alpha$, that f_1 and f_2 are β-th order smooth with the constant $M_\beta \leq \frac{2\alpha^{\frac{1}{\beta}}\beta}{\theta}$ and $(4\mu - \frac{3}{2}\frac{\alpha}{\theta^2})$-strongly convex, and that $f_i(0) - f_i^* \geq \frac{\alpha}{16\mu\theta^2}$ as soon as $\frac{\alpha}{\theta^2} \leq 2\mu$.

Given fixed values for the parameters β and M, the choices of $\alpha = T^{-1/2}$ and $\theta = cT^{-1/2\beta}$ where $c = \frac{2\beta}{M}$ ensure that $\alpha/\theta^2 \leq 2\mu$ as soon as $T \geq (2\mu c^2)^{-\frac{2\beta}{\beta-2}}$ and that the mappings f_1 and f_2 are μ-strongly convex and β-th order smooth with a constant $M_\beta \leq M$.

Moreover, since $\|f_1 - f_2\|_\infty \leq 1/\sqrt{T}$, f_1 and f_2 are undistinguishable with only T queries and thus any algorithm must suffer, when facing f_1 or f_2 an error of the order of

$$\min_x \max\left\{f_1(x) - f_1^*, f_2(x) - f_2^*\right\} = f_1(0) - f_1^* \geq \frac{1}{64}\frac{M}{\mu\beta^2}T^{-\frac{\beta-1}{\beta}}.$$

6. Online Optimization

In the online optimization setting, we have to modify algorithms that use non-uniform averaging as the regret is computed with respect to the Cesaro average of the losses. The online version of the

algorithms are described in Eq. (7) and Eq. (8). The difference with the algorithms of the stochastic case is simply that f is replaced by f_n.

For the two-point algorithm, we recall that it requires that each loss functions can be queried twice, but we emphasize again that the noise is different for the two evaluations and do not cancel simply by differencing.

$$x_n = x_{n-1} - \gamma_n \frac{d}{2\delta_n}\big[f_n(x_{n-1} + \delta_n r_n u_n) - f_n(x_{n-1} - \delta_n r_n u_n) + \varepsilon_n\big] k(r_n) u_n, \tag{7}$$

where γ_n and δ_n depend on n.

The 1-point algorithm evaluates once each loss function and rewrites as

$$x_n = \Pi_K\Big(x_{n-1} - \gamma_n \frac{d}{\delta_n}\big[f_n(x_{n-1} + \delta_n r_n u_n) + \varepsilon_n\big] k(r_n) u_n\Big), \tag{8}$$

where the parameters γ_n and δ_n can evolve with time.

Proposition 9 *Assume each f_n is β-order smooth and M_1-Lipschitz. Then the online version of the algorithms described in Eq. (8) and Eq. (7), with adapted choices of parameters, ensures the following upper-bound on the regret $\frac{1}{N}\sum \mathbb{E}\big[f_n(x_{n-1}) - f_n(x)\big]$:*

– *for unconstrained online optimization of convex mapppings,*

$$\Big(\frac{d^2}{N}\Big)^{\frac{\beta-1}{2\beta}}\left(7\beta M_2\|x_0 - x_*\| + 3\sigma + \Big(\frac{M_\beta}{M_2}\Big)^{\frac{2\beta}{\beta+1}} + \frac{\beta}{N^{1/\beta}}\Big(\frac{M_\beta}{M_2}\Big)^{\frac{-\beta}{\beta+1}}\right)^2.$$

– *for unconstrained online optimization of strongly convex mapppings,*

$$2\beta^2\left(\frac{d^2 M_\beta^2}{N\mu}\right)^{\frac{\beta}{\beta+2}}\big(\sigma^2 + 6\big) + 4\beta^2\frac{d^2 M_1^2 \log(N+1)}{N\mu},$$

– *for constrained online optimization of convex mapppings,*

$$25 R M_\beta\left(\frac{d^2\beta}{N}\right)^{\frac{\beta-1}{2\beta}}(C_{\delta_1} + \sigma^2 + 1),$$

– *for constrained online optimization of strongly convex mapppings,*

$$\Big(\frac{d^2 M_\beta^2}{n\mu}\Big)^{\frac{\beta-1}{\beta+1}}\left(8\beta M_\beta\|x_0 - x_*\| + 4\sigma + 2 + \beta\Big(\frac{M_\beta}{M_2}\Big)^2\Big(\frac{M_\beta^2}{n\mu}\Big)^{\frac{2}{\beta+1}}\right)^2.$$

Actually, the proof are identical in the online optimization setting than in stochastic optimization. The main difference is that we do not use the convexity of f to lower-bound $\frac{1}{N}\sum \mathbb{E}\big[f(x_{n-1}) - f(x)\big]$ by $\mathbb{E}\big[f(\overline{x}_N) - f(x)\big]$.

7. Conclusion

In this paper, we have considered zero-th order online optimization with a special focus on highly-smooth functions such as for online logistic regression. We considered one-point estimates and two-point estimates of the gradient (with then two independent noises). For infinitely differentiable functions, our main result leads to the same dependence on sample size as gradient-based algorithms, with an extra dimension-dependent factor.

The present analysis could be extended in a number of ways: (a) we do not cover the bandit setting. A simple extension of our results allows us to recover existing bounds for $\beta = 1$ (Shamir, 2013) but we are currently unable to obtain high-smoothness improvements for $\beta > 1$; (b) while the two-point analysis considers unconstrained problems, the one-point analysis still requires a compact set of constraints and queries slightly outside (in a δ band around it), which might be avoided by using barrier tools like done by Hazan and Levy (2014). Finally, (c) in the strongly-convex case, the dependence on sample size is optimal in the optimization setting (Polyak and Tsybakov, 1990), however, the optimality of the scaling in dimension, of the plain convex case, and beyond the optimization setting remains open.

Acknowledgments

Part of this work was performed when Francis Bach was holding the Schlumberger chair at IHES and when Vianney Perchet was a researcher at INRIA. Vianney Perchet also acknowledges fundings from the ANR under grant number ANR-13-JS01-0004 and the CNRS under grant project Parasol.

References

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proc. COLT*, 2010.

A. Agarwal, D. Foster, D. Hsu, S. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM J. Optim.*, 23:188–212, 2013.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Adv. NIPS*, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Adv. NIPS*, 2013.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

H. Chen. Lower rate of convergence for locating a maximum of a function. *The Annals of Statistics*, 16(3):1330–1334, 1988.

J. Dippon. Accelerated randomized stochastic optimization. *Ann. Statist.*, 31(4):1260–1281, 08 2003.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: the power of two function evaluations. Technical Report 1312.2139, arXiv, 2013.

V. Fabian. Stochastic approximation of minima with improved asymptotic speed. *Ann. Math. Statist.*, 38(1):191–200, 02 1967.

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proc. Symposium on Discrete algorithms (SODA)*. Society for Industrial and Applied Mathematics, 2005.

E. Hazan and K. Levy. Bandit convex optimization: Towards tight bounds. In *Adv. NIPS*, 2014.

E. Hazan, T. Koren, and K. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proc. Conference On Learning Theory (COLT)*, 2014.

C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, 2009.

S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. Technical Report 0911.0054-v2, ArXiv, 2009.

H. Kushner and G G. Yin. *Stochastic approximation and Recursive Algorithms and Applications*, volume 35. Springer, 2003.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical programming*, 134(2):425–458, 2012.

T. Liang, H. Narayanan, and S. Sakhalin. On zeroth-order stochastic convex optimization via random walks. Technical Report 1402.2667, arXiv, 2014.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley & Sons, 1983.

Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture Notes*, 2004.

Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

Y. Nesterov. Random gradient-free minimization of convex functions. Technical report, Université catholique de Louvain (CORE), 2011.

B. T. Polyak and A. B. Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990.

B. Recht, G. G. Jamieson, and R. Nowak. Query complexity of derivative-free optimization. In *Adv. NIPS*, 2012.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

A. Saha and A. Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proc. Conference on Learning Theory*, 2013.

J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010.

Appendix A. Proof of technical lemmas

A.1. Proof of Lemma 1

This result is rather classical and we first recall the proof when f is twice continuously differentiable. By Taylor expansion, for any $x, y \in \mathbb{R}^d$ and $\lambda > 0$, there exists $\zeta_+ \in [x, x+\lambda y]$ and $\zeta_- \in [x, x-\lambda y]$ such that

$$\begin{aligned} f(x+\lambda y) &= f(x) + \lambda \nabla f(x)^\top y + \frac{\lambda^2}{2} y^\top D^2 f(\zeta_+) y \\ f(x-\lambda y) &= f(x) - \lambda \nabla f(x)^\top y + \frac{\lambda^2}{2} y^\top D^2 f(\zeta_-) y, \end{aligned}$$

This implies that

$$\begin{aligned} \nabla f(x)^\top y &= \frac{f(x+\lambda y) - f(x-\lambda y)}{\lambda} + \frac{\lambda}{2} y^\top D^2 f(\zeta_-) y - \frac{\lambda^2}{2} y^\top D^2 f(\zeta_+) y \\ &\leq \frac{M_0}{\lambda} + \lambda M_2^2 \|y\|^2 \leq 2\sqrt{M_0 M_2^2}\|y\|, \end{aligned}$$

and this yields that $M_1 \leq 2\sqrt{M_0 M_2^2}$. The general proof is obtained by introducing β different number $\lambda_1, \dots, \lambda_\beta$, writing the β equations

$$\left| f(x+\lambda_i y) - \sum_{|m| \leq \beta - 1} \frac{\lambda_i^m}{m!} f^{(m)}(x) y^m \right| \leq \frac{\lambda_i^\beta M_\beta^\beta}{\beta!} \|y\|^\beta,$$

and inverting the system (which is possible if λ_i are all distinct).

A.2. Proof of smoothing lemma

The identity in Eq. (4) is a consequence of the result from Nemirovski and Yudin (1983). Using the smoothness assumption, we have for all $x \in \mathbb{R}^d$:

$$\begin{aligned} & \left| \hat{f}_\delta(x) - f(x) \right| \\ \leq & \left| \mathbb{E}_r \mathbb{E}_{\|v\| \leq 1} r k(r) \sum_{1 \leq |m| \leq \beta - 1} \frac{r^{|m|} \delta^{|m|}}{m!} f^{(m)}(x) v^m \right| + \frac{M_\beta^\beta}{\beta!} \delta^\beta \Big(\mathbb{E}_r |k(r) r^{\beta+1}| \Big) \Big(\mathbb{E}_{\|v\| \leq 1} \|v\|^\beta \Big) \\ \leq & \left| \sum_{1 \leq |m| \leq \beta - 1} \frac{\big(\mathbb{E}_r r^{|m|+1} k(r) \big) \delta^{|m|}}{m!} f^{(m)}(x) \mathbb{E}_{\|v\| \leq 1} \big(v^m \big) \right| + \frac{M_\beta^\beta}{\beta!} \delta^\beta \Big(\mathbb{E}_r |k(r) r^{\beta+1}| \Big) \Big(\mathbb{E}_{\|v\| \leq 1} \|v\|^\beta \Big). \end{aligned}$$

For $|m|$ odd, then, by symmetry of the uniform distribution on the unit ball, $\mathbb{E}_{\|v\| \leq 1} \big(v^m \big) = 0$. Therefore, if $\mathbb{E}_r r^k k(r) = 0$ for k odd and $3 \leq k \leq \beta$, we have:

$$\left| \hat{f}_\delta(x) - f(x) \right| \leq \frac{M_\beta^\beta}{\beta!} \delta^\beta \Big(\mathbb{E}_r |k(r) r^{\beta+1}| \Big).$$

In order to prove the following result on gradients

$$\| \hat{f}'_\delta - f'(x) \| \leq \frac{M_\beta^\beta}{(\beta-1)!} \delta^{\beta-1} \Big(\mathbb{E}_r |k(r) r^\beta| \Big),$$

we first assume that the $\beta+1$-th order derivative tensor is bounded, which will be sufficient by a density argument. In this case, as shown by Nemirovski (2004, p. 38), for all x, the β-th order tensor has projections on $\beta-1$ copies of the vector u and a vector v which is less than $M_\beta^\beta \|u\|^{\beta-1}\|v\|$. This implies that we can apply the function value result to the function $g(x) = f'(x)^\top v$, for any u. This leads to the desired result.

A.3. Bounds on function $k(r)$

From the explicit parameter expansion of Legendre polynomials, we have, for any $\alpha \geq 0$,

$$L'_{2\alpha+1}(0) = \frac{(-1)^\alpha(\alpha+1)}{2^{2\alpha}}\binom{2\alpha+1}{\alpha} = \frac{(-1)^\alpha(2\alpha+1)}{2^{2\alpha}}\binom{2\alpha}{\alpha}.$$

Moreover, we use the following bound obtained from bounds on Catalan numbers: $\binom{2\alpha}{\alpha} \leq \frac{4^\alpha}{\sqrt{\pi\alpha}}$. This leads to $|L'_{2\alpha+1}(0)| \leq \frac{2\alpha+1}{\sqrt{\pi\alpha}}$ for $\alpha > 0$, while for $\alpha = 0$, $|L'_{2\alpha+1}(0)| = 1$

Moreover, for $\beta \geq 3$:

$$\begin{aligned}
\mathbb{E}_r|k(r)|^2 &= \sum_{\alpha=0}^{\lfloor(\beta-1)/2\rfloor}(4\alpha+3)|L'_{2\alpha+1}(0)|^2 \leq 3 + \sum_{\alpha=1}^{\lfloor(\beta-1)/2\rfloor}(4\alpha+3)\frac{(2\alpha+1)^2}{\pi\alpha} \\
&\leq 3 + \sum_{\alpha=1}^{\lfloor(\beta-1)/2\rfloor} 7\alpha\frac{(3\alpha)^2}{\pi\alpha} \leq 3 + \frac{63}{\pi}\frac{(\beta/2)(\beta/2+1)(\beta+1)}{6} \\
&\leq 3 + 21\frac{(\beta/2)(\beta/2+\beta/3)(\beta+\beta/3)}{6} = 3 + \beta^3\frac{21\times 5\times 4}{36\times 12} \leq 3\beta^3.
\end{aligned}$$

This is trivially valid for $\beta = 1$ and $\beta = 2$.

Finally, we have for $\beta \geq 3$:

$$\begin{aligned}
\mathbb{E}_r|k(r)|^2 r^2 &= \sum_{\alpha,\alpha'=0}^{\lfloor(\beta-1)/2\rfloor} \sqrt{4\alpha+3}\sqrt{4\alpha'+3}\mathbb{E}_r\big[L_{2\alpha+1}(r)L_{2\alpha'+1}(r)r^2\big]L'_{2\alpha+1}(0)L'_{2\alpha'+1}(0) \\
&= \sum_{\alpha,\alpha'=0}^{\lfloor(\beta-1)/2\rfloor} \sqrt{4\alpha+3}\sqrt{4\alpha'+3}L'_{2\alpha+1}(0)L'_{2\alpha'+1}(0) \\
&\quad \times\mathbb{E}_r\left[\frac{\big[(2\alpha+2)L_{2\alpha+2}(r)+(2\alpha+1)L_{2\alpha}(r)\big]}{4\alpha+3}\right]\left[\frac{\big[(2\alpha'+2)L_{2\alpha'+2}(r)+(2\alpha'+1)L_{2\alpha'}(r)\big]}{4\alpha'+3}\right] \\
&= \sum_{\alpha,\alpha'=0}^{\lfloor(\beta-1)/2\rfloor} \big(\sqrt{4\alpha+3}\sqrt{4\alpha'+3}\big)^{-1}L'_{2\alpha+1}(0)L'_{2\alpha'+1}(0) \\
&\quad \times\Big[\big[(2\alpha+2)^2+(2\alpha+1)^2\big]\delta_{\alpha=\alpha'}+(2\alpha+1)2\alpha\delta_{\alpha=\alpha'+1}+(2\alpha'+1)2\alpha'\delta_{\alpha'=\alpha+1}\Big] \\
&= \sum_{\alpha=0}^{\lfloor(\beta-1)/2\rfloor} (4\alpha+3)^{-1}L'_{2\alpha+1}(0)^2\big[(2\alpha+2)^2+(2\alpha+1)^2\big] \\
&\quad +2\sum_{\alpha=0}^{\lfloor(\beta-1)/2\rfloor-1} \big(\sqrt{4\alpha+3}\sqrt{4\alpha+7}\big)^{-1}L'_{2\alpha+1}(0)L'_{2\alpha+3}(0)2\alpha(2\alpha+1) \\
&= \sum_{\alpha=0}^{\lfloor(\beta-1)/2\rfloor} (4\alpha+3)^{-1}L'_{2\alpha+1}(0)^2\big[(2\alpha+2)^2+(2\alpha+1)^2\big] \\
&\quad -2\sum_{\alpha=0}^{\lfloor(\beta-1)/2\rfloor-1} \big(\sqrt{4\alpha+3}\sqrt{4\alpha+7}\big)^{-1}L'_{2\alpha+1}(0)^2\frac{(2\alpha+3)(2\alpha+2)}{4(\alpha+1)^2}2\alpha(2\alpha+1) \\
&\leq (4\alpha+3)^{-1}L'_{2\alpha+1}(0)^2\big[8\alpha^2+12\alpha+5\big]\Big|_{\alpha=\lfloor(\beta-1)/2\rfloor} \\
&\quad +\sum_{\alpha=0}^{\lfloor(\beta-1)/2\rfloor-1} (4\alpha+3)^{-1}L'_{2\alpha+1}(0)^2\Big[8\alpha^2+12\alpha+5-(2\alpha+1)4\alpha\Big] \\
&\leq (4\alpha+3)^{-1}\big[\frac{(2\alpha+1)^2}{\pi\alpha}\big]\big]\big[8\alpha^2+12\alpha+5\big]\Big|_{\alpha=\lfloor(\beta-1)/2\rfloor} \\
&\quad +\frac{5}{3}L'_1(0)+\sum_{\alpha=1}^{\lfloor(\beta-1)/2\rfloor-1} (4\alpha+3)^{-1}\big[\frac{(2\alpha+1)^2}{\pi\alpha}\big]\big]\Big[8\alpha^2+12\alpha+5-(2\alpha+1)4\alpha\Big] \\
&\leq (2\beta)^{-1}\big[\frac{\beta^2}{\pi(\beta-2)/2}\big]\big]\big[8\beta^2/4+12\beta/2+5\big] \\
&\quad +\frac{5}{3}+\sum_{\alpha=1}^{\lfloor(\beta-1)/2\rfloor-1} (4\alpha)^{-1}\big[\frac{9\alpha^2}{\pi\alpha}\big]9\alpha \\
&\leq \big[\frac{\beta}{\pi(\beta-2)}\big]\big]\big[2\beta^2+6\beta+5\big]+\frac{5}{3}+\frac{81}{4\pi}\frac{\beta}{2}(\beta/2+1) \leq 2\beta^2+2\beta^2+5+5/3+\frac{81}{16\pi}\beta^2\frac{5}{3} \leq 8\beta^2
\end{aligned}$$

using the three-term recursion formula for Legendre polynomials.

We will also need the following bounds:

$$\begin{aligned}\mathbb{E}_r|k(r)r^{\beta+1}| &\leq \sqrt{\mathbb{E}_r|k(r)^2r^{2\beta+2}|} = 2\sqrt{2}\beta,\\ \mathbb{E}_r|k(r)|^2r^{2+\gamma} &\leq \mathbb{E}_r|k(r)|^2r^2 \leq 8\beta^2 \text{ for any } \gamma \geq 0.\end{aligned}$$

Appendix B. Analysis of classic Stochastic Gradient Descents algorithms

We recall in this section the classical proofs of stochastic gradient descents (see, e.g. Bubeck, 2015, and references therein). We first start when the mappings f_n are not necessarily μ-strongly convex.

Proposition 10 (SGD non-strongly convex) *The stochastic gradient descent*

$$x_n = \Pi_K(x_n - \gamma_n g_n) \tag{9}$$

where g_n is a biased estimate of $f_n'(x_{n-1})$, i.e., such that $\mathbb{E}[g_n|\mathcal{F}_{n-1}] = f_n'(x_{n-1}) + \zeta_n$, and γ_n is non-decreasing achieves the following guarantee

$$\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[f_n(x_{n-1}) - f_n(x)\big] \leq \frac{\max_n \mathbb{E}\|x_n - x\|^2}{2\gamma_N} + \frac{1}{N}\sum_{n=1}^N \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{N}\sum_{n=1}^N \gamma_n^2\mathbb{E}\|g_n\|^2\,.$$

In particular, if $f_n = f$ and $x^\star$ is a minimizer of f, we obtain

$$\mathbb{E}f(\overline{x}_{N-1}) - f(x^\star) \leq \frac{\max_n \mathbb{E}\|x_n - x\|^2}{2N\gamma_N} + \frac{1}{N}\sum_{n=1}^N \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{2N}\sum_{n=1}^N \gamma_n\mathbb{E}\|g_n\|^2$$

Proof We have for any $x \in K$, since projecting reduces distances,

$$\begin{aligned}\|x_n - x\|^2 &\leq \|x_{n-1} - x\|^2 - 2\gamma_n g_n + \gamma_n^2\|g_n\|^2\\ \mathbb{E}\|x_n - x\|^2 &\leq \mathbb{E}\|x_{n-1} - x\|^2 - 2\gamma_n\mathbb{E}f_n'(x_{n-1})^\top(x_{n-1} - x) + 2\gamma_n\mathbb{E}\zeta_n^\top(x_{n-1} - x) + \gamma_n^2\mathbb{E}\|g_n\|^2\\ &\leq \mathbb{E}\|x_{n-1} - x\|^2 - 2\gamma_n\mathbb{E}\big[f_n(x_{n-1}) - f_n(x)\big] + 2\gamma_n\mathbb{E}\zeta_n^\top(x_{n-1} - x) + \gamma_n^2\mathbb{E}\|g_n\|^2\,.\end{aligned}$$

This leads to

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[f_n(x_{n-1}) - f_n(x)\big] &\leq \frac{1}{N}\sum_{n=1}^N \frac{\mathbb{E}\|x_{n-1} - x\|^2 - \mathbb{E}\|x_n - x\|^2}{2\gamma_n}\\ &+ \frac{1}{N}\sum_{n=1}^N \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{2N}\sum_{n=1}^N \gamma_n\mathbb{E}\|g_n\|^2\\ &\leq \frac{\max_n \mathbb{E}\|x_n - x\|^2}{2N\gamma_N} + \frac{1}{N}\sum_{n=1}^N \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{2N}\sum_{n=1}^N \gamma_n\mathbb{E}\|g_n\|^2\,.\end{aligned}$$

■

When the mappings f_n are μ-strongly convex, rates are improved as claimed by the following proposition.

Proposition 11 (SGD μ-strongly convex) *The stochastic gradient descent*

$$x_n = \Pi_K(x_{n-1} - \gamma_n g_n) \tag{10}$$

where g_n is a biased estimate of $f_n'(x_{n-1})$, i.e., such that $\mathbb{E}[g_n|\mathcal{F}_{n-1}] = f_n'(x_{n-1}) + \zeta_n$.

– *The choice of $\gamma_n = \frac{1}{\mu n}$ gives*

$$\frac{1}{N}\sum_{n=1}^N \mathbb{E}f_n(x_{n-1}) - \mathbb{E}f_n(x) + \frac{\mu}{2}\|x_N - x\|^2 \leq \frac{1}{N}\sum_{n=1}^N \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{2N}\sum_{n=1}^N \frac{\mathbb{E}\|g_n\|^2}{\mu n} \tag{11}$$

In particular, if $f_n = f$ and $x^\star$ is a minimizer of f, we obtain

$$\mathbb{E}f(\bar{x}_{N-1}) - f(x^\star) + \frac{\mu}{2}\|x_N - x^\star\|^2 \leq \frac{1}{N}\sum_{n=1}^N \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{2N}\sum_{n=1}^N \frac{\mathbb{E}\|g_n\|^2}{\mu n}\,.$$

– *The choice of $\gamma_n = \frac{2}{\mu(n+1)}$ gives*

$$\mathbb{E}f(\hat{x}_{N-1}) - f(x^\star) + \mathbb{E}\|x_n - x\|^2\frac{\mu}{2} \leq \frac{2}{N(N+1)}\sum_{n=1}^N \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{\mu(n+1)}\mathbb{E}\|g_n\|^2, \tag{12}$$

where $\hat{x}_{N-1} = \frac{2}{N(N+1)}\sum_{n=1}^N n x_{n-1}$.

Proof We have for any $x \in K$:

$$\begin{aligned}
\|x_n - x\|^2 &\leq \|x_{n-1} - x\|^2 - 2\gamma_n g_n + \gamma_n^2\|g_n\|^2 \\
\mathbb{E}\|x_n - x\|^2 &\leq \mathbb{E}\|x_{n-1} - x\|^2 - 2\gamma_n\mathbb{E}f_n'(x_{n-1})^\top(x_{n-1} - x) + 2\gamma_n\mathbb{E}\zeta_n^\top(x_{n-1} - x) + \gamma_n^2\mathbb{E}\|g_n\|^2 \\
&\leq \mathbb{E}\|x_{n-1} - x\|^2 - 2\gamma_n\mathbb{E}\big[f_n(x_{n-1}) - f_n(x) + \mu\|x_{n-1} - x\|^2\big] \\
&\qquad + 2\gamma_n\mathbb{E}\zeta_n^\top(x_{n-1} - x) + \gamma_n^2\mathbb{E}\|g_n\|^2\,.
\end{aligned}$$

This leads to

$$\mathbb{E}f_n(x_{n-1}) - f_n(x) \leq \mathbb{E}\|x_{n-1} - x\|^2(\frac{1}{2\gamma_n} - \frac{\mu}{2}) - \mathbb{E}\|x_n - x\|^2\frac{1}{2\gamma_n} + \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{\gamma_n}{2}\mathbb{E}\|g_n\|^2.$$

First, we consider uniform averaging, induced by the choice of $\gamma_n = \frac{1}{\mu n}$. Indeed, it gives

$$\mathbb{E}f_n(x_{n-1}) - f_n(x) \leq \mathbb{E}\|x_{n-1} - x\|^2\frac{(n-1)\mu}{2} - \mathbb{E}\|x_n - x\|^2\frac{n\mu}{2} + \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{2\mu n}\mathbb{E}\|g_n\|^2.$$

Summing over n and averaging gives

$$\frac{1}{N}\sum_{n=1}^N \mathbb{E}f_n(x_{n-1}) - f_n(x) + \|x_N - x\|^2\frac{N\mu}{2} \leq \frac{1}{N}\sum_{n=1}^N \mathbb{E}\zeta_n^\top(x_{n-1} - x) + \frac{1}{2N}\sum_{n=1}^N \frac{\mathbb{E}\|g_n\|^2}{\mu n}\,.$$

We now consider non-uniform averaging when $f_n = f$, induced by the choice of $\gamma_n = \frac{2}{\mu(n+1)}$, which gives

$$\mathbb{E}f(x_{n-1}) - f(x) \leq \mathbb{E}\|x_{n-1} - x\|^2 \frac{(n-1)\mu}{4} - \mathbb{E}\|x_n - x\|^2 \frac{(n+1)\mu}{4} + \mathbb{E}\zeta_n^\top (x_{n-1} - x) + \frac{1}{\mu(n+1)} \mathbb{E}\|g_n\|^2$$

Multiplying by n, summing, averaging and using the convexity of f yield

$$\mathbb{E}f(\hat{x}_{N-1}) - f(x^\star) + \mathbb{E}\|x_n - x\|^2 \frac{\mu}{2} \leq \frac{2}{N(N+1)} \sum_{n=1}^{N} \mathbb{E}\zeta_n^\top (x_{n-1} - x) + \frac{1}{\mu(n+1)} \mathbb{E}\|g_n\|^2 .$$

■

Appendix C. Proof of Propositions for Unconstrained Optimization

C.1. Proof of Proposition 3

Our iteration is

$$x_n = x_{n-1} - \gamma_n \frac{d}{2\delta_n} \big[f(x_{n-1} + \delta_n r_n u_n) - f(x_{n-1} - \delta_n r_n u_n) + \varepsilon_n \big] k(r_n) u_n .$$

We consider

$$g_n \quad = \quad \frac{d}{2\delta_n} \big[f(x_{n-1} + \delta_n r_n u_n) - f(x_{n-1} - \delta_n r_n u_n) \big] k(r_n) u_n .$$

We will need the expansion using the β-th order smoothness as:

$$f(x_{n-1} + \delta_n r_n u_n) - f(x_{n-1} - \delta_n r_n u_n) = \sum_{|m| \leqslant \beta - 1} \frac{1}{m!} f^{(m)}(x_{n-1}) \big[(\delta_n r_n)^m - (-\delta_n r_n)^m \big] + [A'_n - B'_n],$$

with $|A'_n|, |B'_n| \leqslant \frac{M_\beta^\beta}{\beta!} \delta_n^\beta r_n^\beta$. When taking expectations above, we get exactly the term $2\delta_n f'(x_{n-1})^\top u_n$.

Moreover, since f is 2-smooth

$$\begin{aligned} \big| f(x_{n-1} + \delta_n r_n u_n) - f(x_{n-1} - \delta_n r_n u_n) \big| &\leq M_2^2 r_n^2 \delta_n^2 + 2 |f'(x_{n-1})^\top (\delta_n r_n u_n)| \\ &\leq M_2^2 r_n^2 \delta_n^2 + 2\delta_n r_n |f'(x_{n-1})^\top u_n| . \end{aligned}$$

We then get:

$$\begin{aligned} \mathbb{E}(\|g_n\|^2 | \mathcal{F}_{n-1}) &\leq \frac{d^2 \sigma^2}{4\delta^2} \mathbb{E}_r \big[k(r)^2 \big] + \frac{d^2}{4\delta^2} 2 M_2^4 \delta^4 \mathbb{E}[r^4 k(r)^2] + \frac{d^2}{4\delta^2} 8\delta^2 \mathbb{E}[r^2 k(r)^2] \mathbb{E}\big[|f'(x_{n-1})^\top u_n|^2 \big| \mathcal{F}_{n-1} \big] \\ &\leq \frac{3\beta^3 d^2 \sigma^2}{4\delta_n^2} + 4 d^2 \beta^2 M_2^4 \delta_n^2 + 16 d \beta^2 \mathbb{E}\big[\|f'(x_{n-1})\|^2 \mathcal{F}_{n-1} \big] \text{ using } \mathbb{E} u_n u_n^\top = \frac{1}{d} I , \\ &\leq \frac{3\beta^3 d^2 \sigma^2}{4\delta_n^2} + 4 d^2 \beta^2 M_2^4 \delta_n^2 + 12 d M_2^2 \beta^2 \big[f(x_{n-1}) - f(x_*) \big] , \end{aligned}$$

where we used that $\|f'(x_{n-1})\|^2 \leq 2M_2^2 \big[f(x_{n-1}) - f(x_*) \big]$ for x_* a global optimizer of f.

Thus,

$$\begin{aligned}
&\|x_n - x\|^2 \\
= \;& \|x_{n-1} - x\|^2 - 2\gamma_n (x_{n-1} - x)^\top \big[g_n + \frac{d}{2\delta_n} \varepsilon_n k(r_n) u_n \big] + \gamma_n^2 \Big\| g_n + \frac{d}{2\delta_n} \varepsilon_n k(r_n) u_n \Big\|^2 \\
= \;& \|x_{n-1} - x\|^2 - 2\gamma_n (x_{n-1} - x)^\top \big[g_n + \frac{d}{2\delta_n} \varepsilon_n k(r_n) u_n \big] + 2\gamma_n^2 \|g_n\|^2 + 2\gamma_n^2 \Big\| \frac{d}{2\delta_n} \varepsilon_n k(r_n) u_n \Big\|^2 .
\end{aligned}$$

By taking conditional expectations, we get, using $\mathbb{E} d r_n k(r_n) u_n u_n^\top = I$, and the fact that the expectation of all powers $r_n^\alpha k(r_n)$, $\alpha > 1$, lead to zero:

$$\begin{aligned}
&\mathbb{E}\big[\|x_n - x\|^2 | \mathcal{F}_{n-1} \big] \\
\leqslant \;& \|x_{n-1} - x\|^2 - 2\gamma_n (x_{n-1} - x)^\top f'(x_{n-1}) + 2\gamma_n \mathbb{E} \| \frac{d}{2\delta_n} [A_n' - B_n'] k(r_n) u_n \| \|x_{n-1} - x\| \\
& + 2\gamma_n^2 \mathbb{E}(\|g_n\|^2 | \mathcal{F}_{n-1}) + 2\gamma_n^2 \mathbb{E} \Big\| \frac{d}{2\delta_n} \varepsilon_n k(r_n) u_n \Big\|^2 \\
\leqslant \;& \|x_{n-1} - x\|^2 - 2\gamma_n \big[f(x_{n-1}) - f(x) \big] + \gamma_n d \mathbb{E} \| \frac{1}{\delta_n} \frac{M_\beta^\beta}{\beta!} \delta_n^\beta r_n^\beta k(r_n) u_n \| \|x_{n-1} - x\| \\
& + 2\gamma_n^2 \Big[\frac{3\beta^3 d^2 \sigma^2}{4\delta_n^2} + 4 d^2 \beta^2 M_2^4 \delta_n^2 + 12 d M_2^2 \beta^2 \big[f(x_{n-1}) - f(x_*) \big] \Big] + 2\gamma_n^2 \big(\frac{d}{2\delta_n} \big)^2 \sigma^2 \mathbb{E} k(r_n)^2 \\
\leqslant \;& \|x_{n-1} - x\|^2 - 2\gamma_n \big[f(x_{n-1}) - f(x) \big] + \gamma_n d \delta_n^{\beta - 1} \frac{M_\beta^\beta}{\beta!} 2\beta^2 \|x_{n-1} - x\| \\
& + 2\gamma_n^2 \Big[\frac{3\beta^3 d^2 \sigma^2}{4\delta_n^2} + 4 d^2 \beta^2 M_2^4 \delta_n^2 + 12 d M_2^2 \beta^2 \big[f(x_{n-1}) - f(x_*) \big] \Big] + 6\gamma_n^2 \big(\frac{d}{2\delta_n} \big)^2 \sigma^2 \beta^3 .
\end{aligned}$$

For simplicity, we assume that $\gamma_n = \gamma$ is constant and less than $\frac{1}{24 d M_2^2 \beta^2}$, and that $\delta_n = \delta$. We thus get, with $x = x_*$:

$$\begin{aligned}
\mathbb{E} f(x_{n-1}) - f(x_*) \;\leqslant\; & \frac{1}{\gamma} \mathbb{E} \|x_{n-1} - x_*\|^2 - \frac{1}{\gamma} \mathbb{E} \|x_n - x_*\|^2 \\
& + 2\gamma \Big[\frac{3\beta^3 d^2 \sigma^2}{4\delta^2} + 4 d^2 \beta^2 M_2^4 \delta^2 \Big] + 6\gamma \big(\frac{d}{2\delta} \big)^2 \sigma^2 \beta^3 + d \delta^{\beta - 1} \frac{M_\beta^\beta}{\beta!} 2\beta^2 \sqrt{\mathbb{E} \|x_{n-1} - x_*\|^2} .
\end{aligned}$$

Thus

$$\begin{aligned}
\sum_{n=1}^N \mathbb{E}\big[f(x_{n-1}) - f(x) \big] + \frac{1}{\gamma} \mathbb{E} \|x_N - x_*\|^2 \;\leqslant\; & \frac{1}{\gamma} \|x_0 - x_*\|^2 + 3N\gamma d^2 \sigma^2 \delta^{-2} \beta^3 + 8N\gamma d^2 \beta^2 M_2^4 \delta^2 \\
& + \sum_{n=1}^N 2 d \delta^{\beta - 1} \frac{M_\beta^\beta}{\beta!} \beta^2 \sqrt{\mathbb{E} \|x_{n-1} - x_*\|^2} ,
\end{aligned}$$

which we can put as:

$$\sum_{n=1}^N \big[\mathbb{E} f(x_{n-1}) - f(x_*) \big] + \frac{1}{\gamma} \mathbb{E} \|x_N - x_*\|^2 \leqslant \frac{1}{\gamma} \|x_0 - x_*\|^2 + NC + \sum_{n=1}^N 2 d \delta^{\beta - 1} \frac{M_\beta^\beta}{\beta!} \beta^2 \sqrt{\mathbb{E} \|x_{n-1} - x_*\|^2} ,$$

with $C = 3\gamma d^2\sigma^2\delta^{-2}\beta^3 + 8\gamma d^2\beta^2 M_2^4\delta^2$. This leads to, with $u_n = \sqrt{\mathbb{E}\|x_n - x_*\|^2}$:

$$u_N^2 \leqslant u_0^2 + \gamma NC + \sum_{n=1}^{N} 2\gamma d\delta^{\beta-1}\frac{M_\beta^\beta}{\beta!}\beta^2 u_n.$$

From Lemma 1 of Schmidt et al. (2011), we get:

$$\begin{aligned} u_N &\leqslant \frac{N}{2}2\gamma d\delta^{\beta-1}\frac{M_\beta^\beta}{\beta!}\beta^2 + \left(u_0^2 + \gamma NC + \big[\frac{N}{2}2\gamma d\delta^{\beta-1}\frac{M_\beta^\beta}{\beta!}\beta^2\big]^2\right)^{1/2} \\ &\leqslant N2\gamma d\delta^{\beta-1}\frac{M_\beta^\beta}{\beta!}\beta^2 + u_0 + (\gamma NC)^{1/2}. \end{aligned}$$

Thus

$$\begin{aligned} &\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}f(x_{n-1}) - f(x_*) \\ \leqslant\ &\frac{1}{\gamma N}\|x_0 - x_*\|^2 + C + D\left(N\gamma D + u_0 + (\gamma NC)^{1/2}\right) \end{aligned}$$

with $D = 2d\delta^{\beta-1}\frac{M_\beta^\beta}{\beta!}\beta^2$.

By setting $\gamma = \dfrac{1}{24dM_2^2\beta^2 N^{(\beta+1)/(2\beta)}}$, and $\delta = \dfrac{\beta}{N^{1/(2\beta)}}(M_\beta^\beta M_2)^{-1/(\beta+1)}$, we get:

$$\begin{aligned} C &\leqslant \frac{d^2}{N^{(\beta+1)/(2\beta)}24dM_2^2\beta^2}\big[3\sigma^2\beta^3\frac{N^{1/\beta}}{\beta^2}(M_\beta^\beta M_2)^{2/(\beta+1)} + 8\beta^2 M_2^4\frac{\beta^2}{N^{1/\beta}}(M_\beta^\beta M_2)^{-2/(\beta+1)}\big] \\ &\leqslant \frac{d}{N^{(\beta-1)/(2\beta)}24\beta}(M_\beta/M_2)^{2\beta/(\beta+1)}\big[3\sigma^2 + 8\frac{\beta^3}{N^{2/\beta}}(M_\beta/M_2)^{-4\beta/(\beta+1)}\big] \\ \frac{1}{\gamma N}\|x_0 - x_*\|^2 &\leqslant \frac{24d\beta^2}{N^{(\beta-1)/(2\beta)}}(M_2\|x_0 - x_*\|)^2 \\ D &\leqslant 2d\frac{M_\beta^\beta}{\beta!}\beta^2\frac{\beta^{\beta-1}}{N^{(\beta-1)/(2\beta)}}(M_\beta^\beta M_2)^{-(\beta-1)/(\beta+1)} \\ &\leqslant 2d\frac{\beta}{N^{(\beta-1)/(2\beta)}}(M_\beta/M_2)^{(\beta-1)/(\beta+1)}M_\beta. \end{aligned}$$

This leads to an overall rate of

$$
\begin{aligned}
& \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}f(x_{n-1}) - f(x) \\
\leqslant\ & 2D^2\gamma N + \frac{2}{\gamma N}\|x_0 - x_*\|^2 + 2C \\
\leqslant\ & 2\left(2d\frac{\beta}{N^{(\beta-1)/(2\beta)}}(M_\beta/M_2)^{(\beta-1)/(\beta+1)}M_\beta\right)^2\frac{1}{24dM_2^2\beta^2N^{(\beta+1)/(2\beta)}}N + \frac{48d\beta^2}{N^{(\beta-1)/(2\beta)}}(M_2\|x_0-x_*\|)^2 \\
& + \frac{2d}{N^{(\beta-1)/(2\beta)}24\beta}(M_\beta/M_2)^{2\beta/(\beta+1)}\big[3\sigma^2 + 8\frac{\beta^3}{N^{2/\beta}}(M_\beta/M_2)^{-4\beta/(\beta+1)}\big] \\
\leqslant\ & \frac{d}{N^{(\beta-1)/(2\beta)}}\left(48\beta^2(M_2\|x_0-x_*\|)^2 + 6\sigma^2(M_\beta/M_2)^{2\beta/(\beta+1)} + \frac{1}{3}(M_\beta/M_2)^{4\beta/(\beta+1)}\right) \\
& \frac{16d}{N^{(\beta-1)/(2\beta)}24}\frac{\beta^2}{N^{2/\beta}}(M_\beta/M_2)^{-2\beta/(\beta+1)} \\
\leqslant\ & \frac{d}{N^{(\beta-1)/(2\beta)}}\left(7\beta M_2\|x_0-x_*\| + 3\sigma + (M_\beta/M_2)^{2\beta/(\beta+1)} + \frac{\beta}{N^{1/\beta}}(M_\beta/M_2)^{-\beta/(\beta+1)}\right)^2,
\end{aligned}
$$

which is almost the desired bound, except the dependence on d, which is in d instead of $d^{(\beta-1)/\beta}$. Like in the proof for constrained optimization, we can choose γ and δ with slightly different scalings in d, that is, $\gamma = \frac{1}{24d^{(\beta-1)/\beta}M_2^2\beta^2N^{(\beta+1)/(2\beta)}}$, and $\delta = \frac{\beta d^{1/\beta}}{N^{1/(2\beta)}}(M_\beta^\beta M_2)^{-1/(\beta+1)}$. The value of γ does not satisfy our constraint when $d^{-1/\beta}N^{(\beta+1)/(2\beta)}$ is less than one, which happens only when the final bound is trivial. Thus, we can safely consider the step-size γ above.

Proof for anytime algorithm By setting $\gamma_n = \frac{1}{24dM_2^2\beta^2n^{(\beta+1)/(2\beta)}}$, and $\delta_n = \frac{\beta}{n^{1/(2\beta)}}(M_\beta^\beta M_2)^{-1/(\beta+1)}$, as a function of n, we obtain an anytime algorithm. In order to analyze it, we can simply recycle the proof techniques of Bach and Moulines (2011) (in particular Abel's summation formula). All sums of the forms $\sum_{n=1}^N n^{-\delta}$ may then be bounded thrtough $\frac{N^{1-\delta}}{1-\delta}$ for $\delta \in (0,1)$ and less than $\frac{1}{\delta-1}$ for $\delta > 1$, with $\sum_{n=1}^N \frac{1}{n} \leq \log(N+1)$. The term $\gamma_n^2\delta_n^{-2}$ leads to an extra factor of $\log(N+1)$ while all other factors only lead to extra *constant* factors which are less than 4. The final bound is thus the same as before up to logarithmic terms

C.2. Proof of Proposition 4

The proof technique is the same as for Proposition 3 in Appendix C.1. The first line that differs is the following, where μ-strong convexity is used:

$$
\begin{aligned}
& \mathbb{E}\big[\|x_n - x_*\|^2|\mathcal{F}_{n-1}\big] \\
\leqslant\ & (1-\mu\gamma_n)\|x_{n-1}-x_*\|^2 - 2\gamma_n\big[f(x_{n-1}) - f(x_*)\big] + \gamma_n d\delta_n^{\beta-1}\frac{M_\beta^\beta}{\beta!}2\beta^2\|x_{n-1}-x\| \\
& + 2\gamma_n^2\left[\frac{3\beta^3d^2\sigma^2}{4\delta_n^2} + 4d^2\beta^2M_2^4\delta_n^2 + 12dM_2^2\beta^2\big[f(x_{n-1}) - f(x_*)\big]\right] + 6\gamma_n^2\big(\frac{d}{2\delta_n}\big)^2\sigma^2\beta^3.
\end{aligned}
$$

If we assume that γ_n is less than $\frac{1}{24dM_2^2\beta^2}$, then we get

$$\begin{aligned}\mathbb{E}f(x_{n-1}) - f(x_*) &\leqslant (\frac{1}{\gamma_n} - \mu)\mathbb{E}\|x_{n-1} - x_*\|^2 - \frac{1}{\gamma_n}\mathbb{E}\|x_n - x_*\|^2 \\ &\quad + 2\gamma_n\left[\frac{3\beta^3 d^2\sigma^2}{2\delta_n^2} + 4d^2\beta^2 M_2^4\delta_n^2\right] + d\delta_n^{\beta-1}\frac{M_\beta^\beta}{\beta!}2\beta^2\sqrt{\mathbb{E}\|x_{n-1} - x_*\|^2}. \end{aligned} \tag{13}$$

In order to bound $\sqrt{\mathbb{E}\|x_{n-1} - x_*\|^2}$, we use the same proof technique than in Appendix C.1, without using strong convexity and from the equation:

$$\begin{aligned}\mathbb{E}\|x_n - x_*\|^2 &\leqslant \mathbb{E}\|x_{n-1} - x_*\|^2 \\ &\quad + 2\gamma_n^2\left[\frac{3\beta^3 d^2\sigma^2}{2\delta_n^2} + 4d^2\beta^2 M_2^4\delta_n^2\right] + \gamma_n d\delta_n^{\beta-1}\frac{M_\beta^\beta}{\beta!}2\beta^2\sqrt{\mathbb{E}\|x_{n-1} - x_*\|^2},\end{aligned}$$

which leads to

$$\begin{aligned}\mathbb{E}\|x_n - x_*\|^2 &\leqslant \mathbb{E}\|x_0 - x_*\|^2 \\ &\quad + 2\sum_{k=1}^n \gamma_k^2\left[\frac{3\beta^3 d^2\sigma^2}{4\delta_k^2} + 4d^2\beta^2 M_2^4\delta_k^2\right] + \sum_{k=1}^n \gamma_k\delta_k^{\beta-1}d\frac{M_\beta^\beta}{\beta!}2\beta^2\sqrt{\mathbb{E}\|x_{k-1} - x_*\|^2} \\ &\leqslant \mathbb{E}\|x_0 - x_*\|^2 \\ &\quad + B + \sum_{k=1}^n \gamma_k\delta_k^{\beta-1}d\frac{M_\beta^\beta}{\beta!}2\beta^2\sqrt{\mathbb{E}\|x_{k-1} - x_*\|^2},\end{aligned}$$

with $B = 2\sum_{k=1}^n \gamma_k^2\left[\frac{3\beta^3 d^2\sigma^2}{4\delta_k^2} + 4d^2\beta^2 M_2^4\delta_k^2\right]$.

Thus, with $u_n = \sqrt{\mathbb{E}\|x_n - x_*\|^2}$, we have:

$$u_n^2 \leqslant u_0^2 + B + \sum_{k=1}^n \gamma_k\delta_k^{\beta-1}d\frac{M_\beta^\beta}{\beta!}2\beta^2 u_k$$

From Lemma 1 of Schmidt et al. (2011), we get:

$$u_n \leqslant \sum_{k=1}^n \gamma_k\delta_k^{\beta-1}d\frac{M_\beta^\beta}{\beta!}2\beta^2 + u_0 + B^{1/2}.$$

We now choose $\gamma_n = \frac{1}{n\mu}$, which is less than $\frac{1}{24dM_2^2\beta^2}$ only for certain values of n (if this is not satisfied, the bound is trivial anyway, so this restriction does not impact the result). We select $\delta_n = \left(\frac{d^2\beta!}{M_\beta^\beta\mu n}\right)^{1/(\beta+1)}$.

Then, we may follow the previous proof and sum Eq. (13), with telescoping elements and the same formulas (except the leading terms in $n\mu$, leading to the following bound:

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^N \mathbb{E}f(x_{n-1}) - f(x) \\ \leqslant\ &(\frac{d^2M_\beta^2}{n\mu})^{(\beta-1)/(\beta+1)}\left(8\beta M_\beta\|x_0 - x_*\| + 4\sigma + 2 + \beta(M_2/M_\beta)^2(\frac{M_\beta^2}{n\mu})^{2/(\beta+1)}\right)^2.\end{aligned}$$

Appendix D. Proof of Propositions in Constrained Optimization

D.1. Proof of Proposition 5

We recall that the gradient estimate is $g_n = \frac{d}{\delta_n}\Big(f(x_{n-1} + \delta_n r_n u_n) + \varepsilon_n\Big)k(r_n)u_n$, so that

$$\mathbb{E}g_n = \hat{f}'_{\delta_n}(x_{n-1}) = f'_n(x_{n-1}) + \zeta_n, \quad \text{with } \|\zeta_n\| \leq 2\sqrt{2}\frac{M_\beta^\beta \beta}{(\beta-1)!}\delta_n^{\beta-1}\,.$$

and the variance of g_n is bounded as

$$\mathbb{E}\|g_n\|^2 \leq 6\beta^3\frac{d^2}{\delta_n^2}(C_{\delta_n}^2 + \sigma^2) \leq 6\beta^3\frac{d^2}{\delta_n^2}(C_{\delta_1}^2 + \sigma^2)$$

Using Proposition 10, along with the specific choices of

$$\gamma_n = \frac{R\delta_n}{\sqrt{\beta^3}d\sqrt{n}} \text{ and } \delta_n^\beta = \frac{d\sqrt{\beta}(\beta-1)!}{\sqrt{n}M_\beta^\beta},$$

lead to

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[f_n(x_{n-1}) - f_n(x)\big] &\leq \frac{R^2}{2\gamma_N N} + 3\beta^3 d^2 \frac{1}{N}\sum_{n=1}^N \frac{\gamma_n}{\delta_n^2}(C_{\delta_n}^2 + \sigma^2) + 2\sqrt{2}\frac{M_\beta^\beta}{(\beta-1)!}\beta R\frac{1}{N}\sum_{n=1}^N \delta_n^{\beta-1} \\
&\leq \frac{R^2}{2\gamma_N N} + \frac{4RM_\beta}{N}\sum_{n=1}^N \left(\frac{d\sqrt{\beta}}{\sqrt{n}}\right)^{\frac{\beta-1}{\beta}}(3C_{\delta_n}^2 + 3\sigma^2 + 2\sqrt{2}) \\
&\leq 25RM_\beta\left(\frac{d^2\beta}{N}\right)^{\frac{\beta-1}{2\beta}}(C_{\delta_1}^2 + \sigma^2 + 1)\,.
\end{aligned}$$

D.2. Proof of Proposition 6

Using the same bounds on the biais and variance of g_n than in the proof of Proposition 5 along with the results of Proposition 11 give

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[f_n(x_{n-1}) - f_n(x)\big] + \frac{\mu}{2}\mathbb{E}\|x_N - x\|^2 &\leq 3\beta^3\frac{d^2}{\mu}\frac{1}{N}\sum_{n=1}^N \frac{1}{n\delta_n^2}(C_{\delta_n}^2 + \sigma^2) \\
&\quad + \frac{2\sqrt{2}\beta M_\beta^\beta}{(\beta-1)!}R\frac{1}{N}\sum_{n=1}^N \delta_n^{\beta-1}
\end{aligned}$$

The specific choice of $\delta_n^{\beta+1} = \frac{\beta d^2 \beta!}{n\mu M_\beta^\beta}$ ensures that the upper bound is smaller than

$$15\beta^2 M_\beta^{\frac{2\beta}{\beta+1}}\left(\frac{d^2}{\mu N}\right)^{\frac{\beta-1}{\beta+1}}(C_{\delta_1} + \sigma^2 + 1)\,.$$

D.3. Proof of Proposition 7

Proposition 7 is another consequence of Propositions 10 and 11. Indeed, for $\beta = 2$, the mapping $\hat{f}_\delta$ is convex, hence we can consider the algorithms as stochastic gradient descents on $\hat{f}_\delta$, with an unbiased estimate of the gradient. Then it suffices to approximate $\hat{f}_\delta(x)$ by $f(x) \pm 2\sqrt{2}\beta\frac{M_\beta^\beta}{\beta!}\delta^\beta$, to choose parameters so that error terms balance and to conclude.

D.4. Proof of Proposition 8

Once again, the proof uses the same standard arguments than the proof of Proposition 11. More precisely, we consider here constant step size $\delta_n = \delta$, where δ is small enough so that $\hat{f}_\delta$ is μ'-strongly convex (where $\mu' \leq \mu$ and, as we will see, it will be implied by N being big enough) and we apply Proposition 11 to $\hat{f}_\delta$, this allows us to bound $\mathbb{E}\|x_N - x^\sharp\|^2$, where $x^\sharp$ is a minimizer of $\hat{f}_\delta$

Finally, we conclude using the smoothness and the strong convexity of f that imply that

$$\|x^\sharp - x^\star\| \leq \frac{1}{\mu'}\|f'(x^\sharp)\| \leq \frac{1}{\mu'}\|f'(x^\sharp) - f'_\delta(x^\sharp)\| \leq \frac{1}{\mu'}\frac{M_\beta^\beta}{(\beta-1)!}\delta^{\beta-1}2\sqrt{2}\beta.$$

As a consequence the triangle inequality

$$\mathbb{E}\|x_N - x^\star\|^2 \quad \leq \quad 2\mathbb{E}\|x_N - x^\sharp\|^2 + 2\|x^\sharp - x^\star\|^2$$

and the combined above majorations of $\mathbb{E}\|x_N - x^\sharp\|^2$ and $2\|x^\sharp - x^\star\|^2$ give the result.

We emphasize agains that the fact that f_δ is μ'-strongly convex is ensured by N being large enough (and the larger N, the bigger $\mu' \leq \mu$ can be chosen).